Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 181-184 recibido 12-04-12 revisado 31-05-12 aceptado 05-06-12

IARG-AnCora: Anotación de los corpus AnCora con argumentos implícitos

IARG-AnCora: Annotating AnCora corpus with implicit arguments

Mariona Taulé, M. Antònia Martí, Aina Peris
CLiC-Universidad de Barcelona
Gran Via 585, 08007 Barcelona
{mtaule,amarti,aperis}@ub.edu

Horacio Rodríguez
TALP-Universidad Politécnica de Cataluña
Jordi Girona Salgado 1-3, Barcelona
horacio@lsi.upc.edu

Lidia Moreno
ELiRF-Universidad Politécnica de Valencia
Camino de Vera s/n, 46020 Valencia
lmoreno@dsic.upv.es

Paloma Moreda
GPLSI-Universidad de Alicante
Campus de San Vicente del Raspeig,
03080, Alicante
moreda@dlsi.ua.es

Resumen: IARG-AnCora tiene como objetivo la anotación con papeles temáticos de los argumentos implícitos de las nominalizaciones deverbales en el corpus AnCora. Estos corpus servirán de base para los sistemas de etiquetado automático de roles semánticos basados en técnicas de aprendizaje automático. Los analizadores semánticos son componentes básicos en las aplicaciones actuales de las tecnologías del lenguaje, en las que se quiere potenciar una comprensión más profunda del texto para realizar inferencias de más alto nivel y obtener así mejoras cualitativas en los resultados.
Palabras clave: Estructura argumental, nominalización deverbal, argumentos implícitos, anotación de corpus, recursos lingüísticos

Abstract: Iarg-AnCora aims to annotate the implicit arguments of deverbal nominalizations in AnCora corpus. This corpus will be the basis for systems of automatic semantic role labeling based on machine learning techniques. Semantic analyzers are essential components in the current applications of language technologies, in which it is important to obtain a deeper understanding of the text to make inferences on the highest level in order to obtain qualitative improvements in the results.
Keywords: Argument structure, deverbal nominalization, implicit arguments, corpus annotation, linguistic resources

1 Motivación y Objetivos

Tradicionalmente el análisis de la estructura argumental se ha centrado principalmente en los predicados verbales, aunque recientemente se ha extendido también a los predicados nominales y adjetivales. En la gran mayoría de estas propuestas la identificación de los argumentos se restringe a los que aparecen en la oración, en el caso de los verbos, y al SN, en el caso de los nombres, es decir a los argumentos explícitos. Lo mismo es aplicable a los sistemas de etiquetado automático de roles semánticos (*Semantic Role Labeling*), que utilizan estos recursos para el aprendizaje del modelo de etiquetado, la mayoría de los cuales se aplican a verbos y sólo reconocen y clasifican los argumentos explícitos (Márquez et al., 2008), (Palmer, Gildea y Xue, 2010). Explicitar esta información permite una cobertura mucho más amplia del contenido semántico de los documentos (Gerber, Chai, y Meyers, 2009).

El proyecto *IARG-AnCora: Anotación de los corpus AnCora con argumentos implícitos*[1]

[1] Acción complementaria (FFI2011-13737-E), asociada al proyecto TextMess 2.0 (TIN2009-13391-C04-03/04).

tiene como objetivo principal enriquecer los corpus AnCora[2] del español y del catalán con la anotación de los argumentos implícitos de los predicados nominales derivados de verbos. Hasta el momento sólo se habían anotado los argumentos explícitos de estos nombres.

AnCora está formado por un corpus del catalán y otro del español de 500.000 palabras cada uno anotado con información morfológica (lema y categoría), sintáctica (constituyentes y funciones), semántica (estructura argumental, roles semánticos, entidades nombradas y sentidos nominales de WordNet) y pragmática (correferencia).

1.1 Argumento implícito

Se entiende por argumento implícito aquel argumento que no se realiza en el sintagma nominal (SN) el núcleo del cual es el nombre deverbal, pero que se encuentra en el contexto oracional (1) o textual (2) de la nominalización.

1) *Las escuelas de samba de Sao Paulo*$_{\textbf{iarg1-pat}}$ han conseguido **[**el **apoyo** *de la empresa privada*$_{\textbf{arg0-agt}}$ para mejorar las fiestas de carnaval**]**.

2) **"***El carnaval de Sao Paulo es feo***"**$_{\textbf{iarg1-pat}}$**,** dijo hoy *el alcalde de Río de Janeiro*$_{\textbf{iarg0-agt}}$ en una conversación informal con periodistas cariocas, y encendió la polémica. **[**Esa **opinión]** fue respaldada por el gobernador de Río de Janeiro, quien incluso fue más allá en su crítica al comentar que el carnaval que se organiza en Sao Paulo es "más aburrido que un desfile militar".

En el ejemplo (1), el nombre deverbal 'apoyo' tiene el argumento agente (arg0-agt) explicitado en el mismo SN, mientras que el argumento paciente 'las escuelas de samba de Sao Paulo' está implícito (iarg1-pat), porque se realiza en la misma oración pero fuera del SN. En el ejemplo (2), en cambio, el nombre deverbal 'opinión' aparece en el SN sin ningún argumento explícito. Sin embargo, tanto el argumento agente, 'el alcalde de Río de Janeiro', como el argumento paciente, 'el carnaval de Sao Paulo es feo', se consideran argumentos implícitos (iarg-agt y iarg-pat, respectivamente) porque se realizan en la oración previa. En la anotación actual del corpus AnCora, sólo los argumentos dentro del SN están anotados, por lo tanto, 'opinión' no tiene ningún argumento asociado y 'apoyo' sólo el argumento agente.

La tarea a desarrollar en este proyecto consiste básicamente en identificar los argumentos implícitos y asignarles una posición argumental –iarg0, iarg1, etc.– con el correspondiente papel temático (agente, paciente, causa, etc.).

Estos argumentos pueden recuperarse si se tiene en cuenta un contexto discursivo más amplio (Ruppenhofer et al., 2009). Tenerlos identificados es, por lo tanto, importante para poder proporcionar una interpretación semántica completa de las oraciones y textos.

2 Corpus con argumentos implícitos anotados

Sólo existen dos corpus que contengan los nombres deverbales anotados con argumentos implícitos y ambos son del inglés:

1) El corpus de entrenamiento y evaluación creado para llevar a cabo la tarea 10 de SemEval-2010, *Linking events and their participants in discourse*[3]. Se trata de un corpus formado por textos literarios de ficción y etiquetado siguiendo el esquema de anotación de FrameNet (Erk y Padó, 2004).
2) Un subconjunto de la sección de entrenamiento, desarrollo y evaluación del corpus periodístico Penn TreeBank (Marcus et al., 1993). El esquema de anotación sigue las propuestas de PropBank (Palmer et al., 2005) y NomBank (Meyers, Reeves, y Macleod, 2004) y (Meyers, 2007).

En el primer caso se han anotado un total de 3.073 ocurrencias que cubren distintos predicados nominales, pero cada uno de ellos asociado con un número pequeño de ocurrencias anotadas. En el segundo caso, sólo se han seleccionado los 10 nombres más frecuentes con sentidos no ambiguos, y se han anotado todas las ocurrencias del subconjunto del corpus en las que aparecen, un total de

[2]Los corpus AnCora están disponibles en la página web: http://clic.ub.edu/corpus/ancora.

[3]http://www.coli.uni-saarland.de/projects/semeval2010_FG/

1.253 (Gerber y Chai, 2010). En ambos corpus sólo se han anotado los argumentos implícitos 'nucleares' de nombres derivados de verbos, en ningún caso los argumentos adjuntos (o periféricos siguiendo la terminología de FrameNet[4]).

IARG-AnCora será el primer corpus anotado semánticamente con argumentos implícitos para el español y catalán. A diferencia de los corpus ingleses, la cobertura de IARG-AnCora será más amplia en dos sentidos: por un lado, se anotarán todos los argumentos implícitos de todas las ocurrencias nominales deverbales de los corpus AnCora (del orden de 23.000 aproximadamente para cada lengua); por el otro lado, se tendrán en cuenta tanto los argumentos implícitos 'nucleares' (iarg0, iarg1, iarg2, iarg3 y iarg4) como los argumentos adjuntos (iargM), entre los que se priorizarán los argumentos locativos (iargM-loc), temporales (iargM-tmp) y finales (iargM-fin).

Estas dos características los diferencian de los corpus anotados con argumentos implícitos.

3 Metodología

Se utilizará el mismo esquema de anotación que se ha adoptado en la anotación de los argumentos explícitos de las nominalizaciones deverbales (Peris y Taulé, 2011), que, a su vez, es el mismo que se ha utilizado para la anotación de la estructura argumental de los verbos (Taulé et al., 2008). El esquema de anotación sigue la propuesta de PropBank y NomBank enriquecida con papeles temáticos. De esta manera, se asegura la consistencia de la anotación de los argumentos entre diferentes predicados ─nombres y verbos─ así como la compatibilidad de los recursos del español y catalán con los del inglés.

En el caso de los argumentos implícitos, la etiqueta que se utilizará es $iarg_n$ para diferenciarlos de los argumentos explícitos (arg_n) (Gerber y Chai, 2010). La lista de papeles temáticos incluye 20 etiquetas distintas[5] ampliamente reconocidas en lingüística. La combinación de las 6 etiquetas argumentales (iarg0, iarg1, iarg2, iarg3, iarg4, iargM) con los distintos papeles temáticos da como resultado un total de 36 etiquetas semánticas posibles (iarg0-cau, iarg1-agt, iarg0-agt, iarg2-loc, etc.) que servirán para describir la relación semántica que se establece entre los argumentos y sus predicados.

La anotación del corpus con argumentos implícitos se realizará en dos etapas, la primera se llevará a cabo de manera automática y la segunda manualmente.

a) En la primera etapa se desarrollará un modelo de etiquetado de roles semánticos basado en técnicas de aprendizaje automático, cuyo objetivo será la identificación y clasificación de los argumentos implícitos y con el cual se etiquetará automáticamente todo el corpus (tanto la parte del español como del catalán). Este modelo se inferirá a partir de un corpus de entrenamiento anotado previamente de manera manual consistente en una muestra seleccionada de 500 ocurrencias nominales del corpus del español.
b) En una segunda etapa se procederá a la revisión manual de la anotación obtenida en el proceso automático anterior con el fin de garantizar la calidad final del recurso. Esta anotación manual servirá también para evaluar la precisión y cobertura del sistema automático desarrollado. Dado que dicho sistema se utilizará para la anotación automática del catalán, será posible analizar también el grado de portabilidad del sistema a otra lengua.

Tanto en el proceso automático como en el manual, se utilizarán los léxicos AnCora-Verb (Aparicio et al., 2008) y AnCora-Nom (Peris y Taulé, 2011) como fuentes léxicas a partir de las cuales obtener información de los argumentos implícitos posibles de cada predicado. Los argumentos potenciales a localizar en el contexto discursivo local, y posteriormente a etiquetar, serán aquellos que aparecen declarados en la entrada léxica nominal o verbal y que no están realizados dentro del SN el núcleo del cual es la nominalización deverbal.

El proyecto dará lugar, por un lado, a una versión enriquecida de los corpus AnCora con información sobre los argumentos implícitos de

[4] http://framenet.icsi.berkeley.edu/

[5] Los papeles temáticos son: 'agt' (agente), 'cau' (causa), 'exp' (experimentador), 'scr' (origen), 'pat' (paciente), 'tem' (tema), 'cot' (cotema), 'atr' (atributo), 'ben' (beneficiario), 'ext' (extensión), 'ins' (instrumento), 'loc' (locativo), 'tmp' (tiempo), 'mnr' (manera), 'ori' (origen), 'des' (destino), 'fin' (finalidad), 'ein' (estado inicial), 'efi' (estado final) y 'adv' (adverbial).

los nombres deverbales y, por otro, se dispondrá de un primer modelo para desarrollar un sistema de etiquetado de roles semánticos basado en rasgos para predicados nominales que contemple todos los argumentos, explícitos e implícitos.

4 Conclusiones

Disponer de corpus de amplia cobertura que incluyan la anotación tanto de los argumentos explícitos como de los implícitos y para predicados verbales y nominales, convierte dichos recursos en una fuente de conocimiento de gran valor. Los corpus AnCora del español y del catalán enriquecidos con dicha información, serán los primeros de esta extensión con este tipo de información. Estos corpus se podrán utilizar tanto para el estudio y análisis de la estructura argumental de nombres y verbos en general, como para derivar sistemas automáticos de etiquetado de roles semánticos que tengan en cuenta tanto los argumentos explícitos como implícitos de los nombres, para el estudio de las cadenas correferenciales, el referente de los SN, etc. La derivación de estos sistemas no se puede llevar a cabo si no se dispone de este tipo de recursos.

Bibliografía

Aparicio, J., M. Taulé y M.A. Martí, (2008). 'AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora'. *Proceedings of 6th International Conference on of Language, Resources and Evaluation*. Marrakech, Morocco.

Erk K. y S. Padó, (2004). 'A powerful and versatile XML Format for representing role-semantic annotation'. *Proceedings of 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal.

Gerber, M., J. Chai, A. Meyers, (2009). 'The role of implicit argumentation in nominal SRL', *Proceedings of Human Language Technologies: NACL-2009*, pp. 146–154, Boulder, Colorado.

Gerber, M. y J.Y. Chai, (2010). 'Beyond NomBank: A Study of Implicit Argumentation for Nominal Predicates. *Proceedings of the ACL conference 2010*, pp. 1583–1592, Uppsala, Sweden, ACL.

Marcus, M., B. Santorini, y M. Marcinkiewicz, (1993). 'Building a large annotated corpus of English: the Penn treebank'. *Computational Linguistics*, 19:313-330.

Márquez, L., X. Carreras, C. Kenneth, Litkowski, y S. Stevenson, (2008). 'Semantic role labeling: an introduction to the special issue'. *Computational Linguistics*, 34(2):145–159.

Meyers, A., R. Reeves, y C. Macleod, (2004). 'NP-external arguments, a study of argument sharing in English. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing (MWE'04)*, pp. 96–103, Stroudsburg, PA, USA. ACL.

Meyers, A. (2007). 'Anotation Guidelines for NomBank-Noun Argument Structure for PropBank'. Technical report, University of New York.

Palmer, M., Kingsbury, P. and Gildea, D. (2005): The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, 21 (1). USA: MIT Press.

Palmer, M., D. Gildea, y N. Xue, (2010). Semantic Role Labeling. Synthesis on Human Languages Technologies. Morgan and Claypool Publishers.

Peris, A y M. Taulé, (2011). 'Annotating the Argument Structure of Deverbal Nominalizations in Spanish'. *Language Resources and Evaluation*, Springer. DOI 10.1007/s10579-011-9172-x.

Peris, A y M. Taulé, (2011). 'AnCora-Nom: A Spanish lexicon of deverbal nominalizations', *Procesamiento del Lenguaje Natural*, nº46, pp. 11-18.

Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C, Palmer, M. (2009). Semeval-2010 task 10: Linking events and their participants in discourse. Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Direvtions (SEW 2009), p. 106-11, ACL, Boulder, Colorado.

Taulé, M., M.A., Martí, y M. Recasens (2008). 'Ancora: Multilevel Annotated Corpora for Catalan and Spanish'. *Proceedings of 6th International Conference on Language Resources and Evaluation.* Marrakesh Morocco.